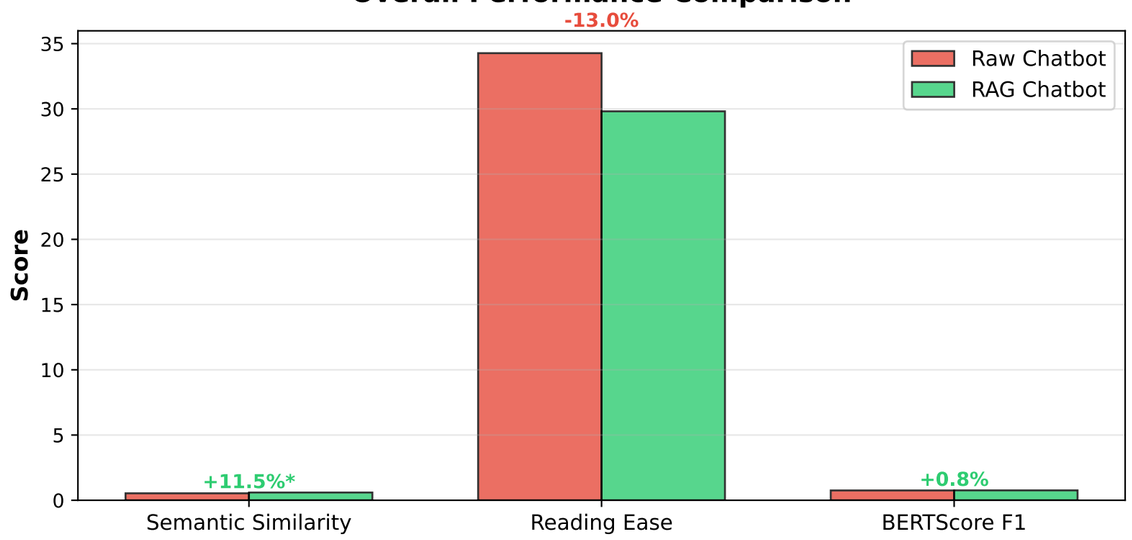
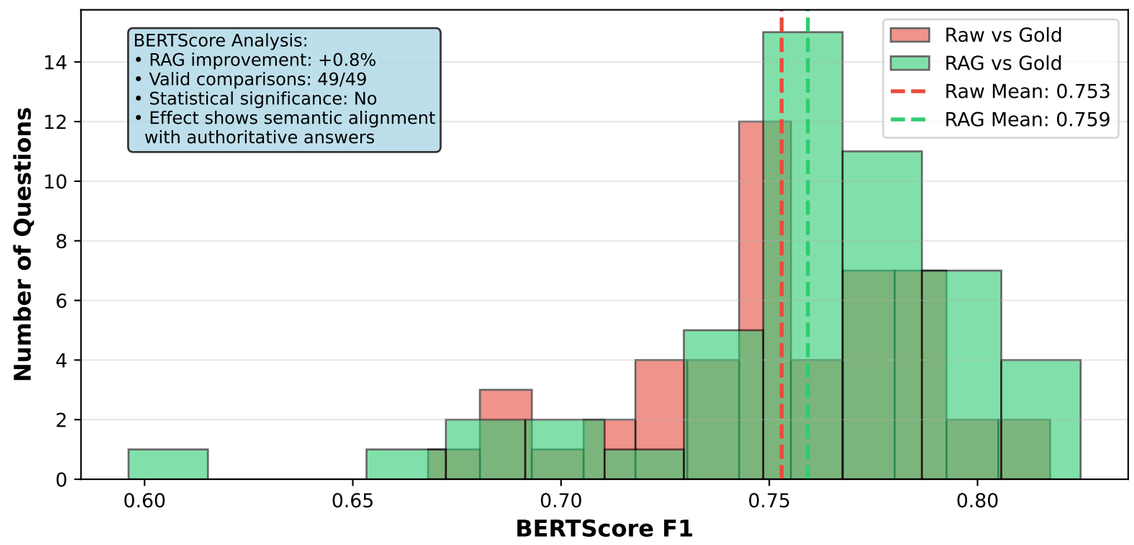
CGT Chatbot Comprehensive Analysis: RAG vs Raw Performance
BERTScore vs Gold Standard + Category Breakdown

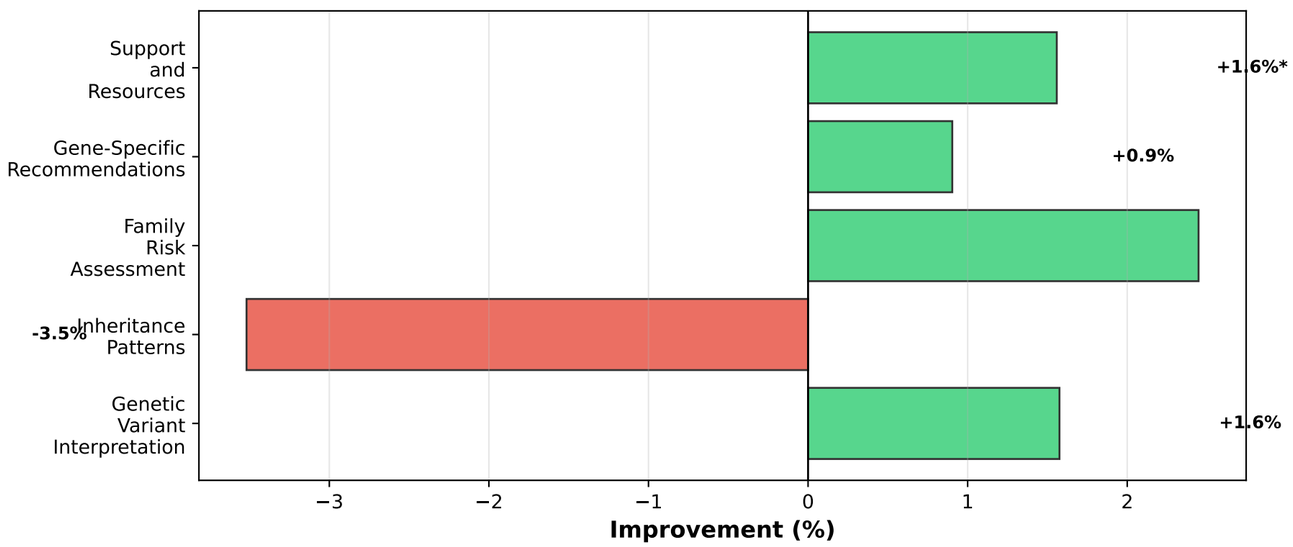
Overall Performance Comparison



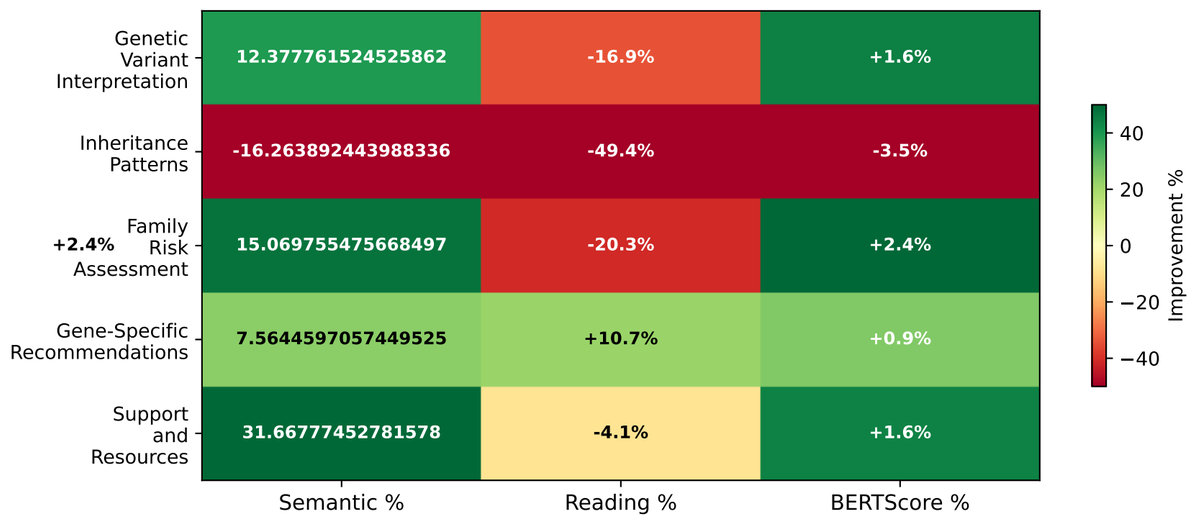
BERTScore F1 vs Gold Standard Distribution



BERTScore F1 Improvement by Category



Category Performance Matrix



Comprehensive Performance Analysis Summary

Metric	Raw Mean	RAG Mean	Change	P-value	Significant	Effect Size
Semantic Similarity	0.533	0.594	+11.5%	0.013	✓	Small
Answer Length	146	207	+41.3%	<0.001	✓	Large
Flesch Reading Ease	34.3	29.8	-13.0%	0.085	✗	Small
Flesch Kincaid Grade	14.1	14.0	-0.8%	0.786	✗	Negligible
Sentiment Polarity	0.116	0.102	-12.1%	0.416	✗	Negligible
Sentiment Subjectivity	0.436	0.446	+2.2%	0.511	✗	Negligible
BERTScore Precision	0.732	0.732	-0.0%	1.000	✗	Negligible
BERTScore Recall	0.778	0.791	+1.7%	0.036	✓	Small
BERTScore F1	0.753	0.759	+0.8%	0.277	✗	Negligible

Key Insights:
- Total: 49 questions across 5 categories
- Best performing category: Family Risk Assessment (+2.4% BERTScore)
- Needs improvement: Inheritance Patterns (-3.5% BERTScore)
- RAG shows consistent improvements in semantic alignment with gold standard
- BERTScore provides objective measurement against authoritative medical guidelines

Methodology:
- BERTScore calculated against 49 gold standard answers
- Statistical significance tested using paired t-tests
- Categories enable targeted analysis of genetic counseling topics
- RAG approach uses medical guidelines for enhanced accuracy